Jamaal Mirville

Google White Paper (Shapley  Summery)

Throughout this whitepaper Google discusses the evolution of machine learning and how to improve it with Explainable AI.  The early phases of machine learning began with heuristics, rules, linear models, decision trees, deep models, ensembles, and now meta-learning to create models. As model development improved, they also improved in areas of expressiveness, versatility, adaptability and efficiency.  Due to improvement model have become more complexed and created new challenges in areas of spurious correlations, loss of debuggability, transparency, proxy objectives loss of control, and undesirable data amplification.  These new issues are believed to be resolved with the help of XAI (eXplainable AI) and Human Reasoning.  While utilizing explanations Google believes it will help build better mental models.

The purposes explanations in AI will inform and support human decision making when AI is being utilized for decision aid. Improving transparency by employing an understanding between AI and humans. Enabling debugging, auditing, verifying generalization ability and moderating trust. With these we should expect the model user to have better decision making and the model builder to improve the model.

Peter Lipton uses the terminology foil to denote why an event did not happen. In XAI if we were to implement and utilize the foil, we could understand parts of the model behavior that we wouldn’t normally understand.  Google has started to offer comprehensive XAI offerings to help meet users needs. These new offerings will help users understand what led the model to its prediction. This offering is in its AI Platform Prediction and in their AutoML Tables. These new offerings will be utilizing the Shapley Additive Explanations.

The idea behind Shapley Additive Explanations is to get an understanding of the importance of a feature and estimate how much that feature has contributed to the model's prediction. This happens by viewing how well the model behaves with and without that feature for every combination of features. This is different from scikit-learn which calculate the global feature importance. This makes it so local features play a role in each data point. The Shapley values informs us how to distribute the prediction among features.

The Shapley value helps satisfy completeness, symmetry, dummy and additivity. Completeness which is also efficiency is supported by the Shapley values summed to the difference between the target outcome and the outcomes of the foil. Google's whitepaper described the axiom for Symmetry as if there are 2 participants i, j and S, where S is any subset of participants not including i, j, when {S, i} = {S, j} → i and j should have the same attribution value. As for Dummy Google's whitepaper also stated that one participant i and S, where S is any subset of participants not including i, when {S, i} = {S} → i should get zero attribution. For Additivity the Shapley values of the sum of the two outcome functions should be the sum of the Shapley values of the two outcome functions.

Due to the usage of foil, the model will now need to employ baselines and counterfactuals. The baseline score model's prediction for the baseline instance. While using

Shapley values we need to utilize a baseline that reflects the desired foil. This will be used to compare against an example form the domain of the model. While doing this we would do this for multiple trials, and this would become our counterfactuals user would rely on the model to predict outcome for these counterfactuals and then will be used to calculate the Shapley values. Due to this we come across two different types of baselines, uninformative and informative. An uninformative is where no additional information is present and the model maps to the most likely foil. Informative involves selecting baselines which have high information content to compare with each instance and highlight the most important feature attribution that will result in the observed outcome difference.

Calculating the Shapley value can become very difficult since it scales exponentially to the number of input features. Google has employed the following methods to help relieve this issue. Integrated gradients which enables computing the Shapley value for differentiable model. Sampled Shapley which provides an approximation through sampling to the discrete Shapley value. Finally, XRAI which is specialized for image inputs and works with integrated gradients with over segmentation and region selection to determine attribution density at level of regions instead of pixels.

Integrated gradients are suggested for neural networks and differentiable models since in offers a better computational advantage for large inputs. Sampled Shapley is recommended for tabular and non-differentiable models.

In conclusion the Explainable AI has brought awareness to issues currently facing machine learning. Although Shapley does help with help with resolving those issues conceptually, but the main problem with sampling every feature will be extremely time consuming. It will with help with the simulation of adding a human in the mix of creating and improving your models.